DNA Research, 2016, 1–10
doi: 10.1093/dnares/dsw013
Full Paper

OXFORD

Full Paper

Diversity and population-genetic properties of copy number variations and multicopy genes in cattle

Derek M. Bickhart[1,†,*], Lingyang Xu[1,2,†], Jana L. Hutchison[1], John B. Cole[1], Daniel J. Null[1], Steven G. Schroeder[1], Jiuzhou Song[2], Jose Fernando Garcia[3], Tad S. Sonstegard[1], Curtis P. Van Tassell[1], Robert D. Schnabel[4,5], Jeremy F. Taylor[4], Harris A. Lewin[6], and George E. Liu[1,*]

[1]USDA-ARS, Animal Genomics and Improvement Laboratory, Beltsville, MD 20705, USA, [2]Department of Animal and Avian Sciences, University of Maryland, College Park, MD 20742, USA, [3]Universidade Estadual Paulista (UNESP), Araçatuba, São Paulo, Brazil, [4]Division of Animal Sciences, University of Missouri, Columbia, MO 65211, USA, [5]Informatics Institute, University of Missouri, Columbia, MO, USA, and [6]Department of Evolution and Ecology, University of California, Davis, CA 95616, USA

*To whom correspondence should be addressed. Tel. +1 301-504-8592. Fax. +1 301-504-8092. E-mail: derek.bickhart@ars.usda.gov (D.M.B.); Tel. +1 301-504-9843. Fax. +1 301-504-8414. E-mail: george.liu@ars.usda.gov (G.E.L.)

[†]These authors are co-first authors.

Edited by Dr Osamu Ohara

Received 5 November 2015; Accepted 29 February 2016

Abstract

The diversity and population genetics of copy number variation (CNV) in domesticated animals are not well understood. In this study, we analysed 75 genomes of major taurine and indicine cattle breeds (including Angus, Brahman, Gir, Holstein, Jersey, Limousin, Nelore, and Romagnola), sequenced to 11-fold coverage to identify 1,853 non-redundant CNV regions. Supported by high validation rates in array comparative genomic hybridization (CGH) and qPCR experiments, these CNV regions accounted for 3.1% (87.5 Mb) of the cattle reference genome, representing a significant increase over previous estimates of the area of the genome that is copy number variable (~2%). Further population genetics and evolutionary genomics analyses based on these CNVs revealed the population structures of the cattle taurine and indicine breeds and uncovered potential diversely selected CNVs near important functional genes, including *AOX1*, *ASZ1*, *GAT*, *GLYAT*, and *KRTAP9-1*. Additionally, 121 CNV gene regions were found to be either breed specific or differentially variable across breeds, such as *RICTOR* in dairy breeds and *PNPLA3* in beef breeds. In contrast, clusters of the *PRP* and *PAG* genes were found to be duplicated in all sequenced animals, suggesting that subfunctionalization, neofunctionalization, or overdominance play roles in diversifying those fertility-related genes. These CNV results provide a new glimpse into the diverse selection histories of cattle breeds and a basis for correlating structural variation with complex traits in the future.

Key words: cattle genome, population sequencing, copy number variation, taurine, indicine

1. Introduction

Copy number variations (CNVs) are deletions and insertions of genomic sequence between two individuals of a species.[1–3] Substantial progress has been made in understanding the impacts of CNVs on both normal phenotypic variability and disease susceptibility in human.[4] The majority of previous studies of CNV in domesticated animals have been based on array comparative genomic hybridization (CGH) experiments or single-nucleotide polymorphism (SNP) arrays.[5–13] As in human,[14] sequence-based approaches are becoming popular for the study of CNVs in domesticated animals.[15–17]

The CNV distribution within and among species seems to be shaped by mutation, selection, and demographic history.[18,19] However, unlike SNPs and microsatellites, the population genetics of CNV is largely unknown.[1,20–23] Several studies have explored the evolution and adaptation aspects of CNVs in human,[24,25] and other species.[26–29] Only a few cases of recent positive selection were found near *AMY1*, *APOBEC*, *MAPT*, *MIF*, and *UGT2B17* within human populations.[30–33] While most of simple deletions and simple duplications (75–90%) display extensive linkage disequilibrium (LD) with SNPs,[1,34,35] the properties of CNVs not tagged by SNPs remain unexplored in cattle. This is mainly due to the difficulty of genotyping CNV and to the limited sample size in the published CNV studies.

As one of the most important farm animals, cattle are used for a variety of purposes including dairy, beef, leather, and labour. The majority of the global cattle population can be classified into one of two subspecies: humpless (taurine, *Bos taurus taurus*) and humped (indicine or zebu, *Bos taurus indicus*) cattle with dramatic phenotypic differences between them.[36,37] Earlier studies indicated that these two subspecies diverged from the last common ancestor between 0.6 and 2 million yrs ago.[38,39] There appears to have been two separate domestication events, with taurine cattle likely being domesticated in the Fertile Crescent ~8,000–10,000 yrs ago and indicine cattle in the Indus Valley ~6,000–8,000 yrs ago.[40,41] A third independent domestication was proposed in Africa;[42,43] however, a recent study did not support this hypothesis.[44] Since the early 1800s, breed development has primarily been based on phenotypic selection on coat colour and polled phenotypes. More recently, the adoption of effective genetic selection programs and widespread use of artificial insemination resulted in bottlenecks followed by breed expansion. During the last 50 yrs, animal breeding based on quantitative genetics has resulted in remarkable progress in improving milk and meat production traits.[45,46] Therefore, selective (natural and human-imposed) and non-selective forces (demographic events and introgression) have driven changes within the cattle genome. Their combined effects have created exceptional phenotypic diversity and genetic adaptation to local environments across the globe within the modern cattle breeds. For example, indicine cattle are better adapted to warm climates and demonstrate superior resistance to tick infestation than do taurine breeds.[47] Likewise, beef and dairy cattle breeds display distinct patterns in selected metabolic pathways related to muscling, marbling, and milk composition traits. Although cattle genome evolution and demographic history have been explored from multiple aspects, the diversity and population genetic properties of CNV in cattle are still unexplored.

In this study, we compare the diversity and population-genetic properties of CNVs in ~70 cattle individuals sequenced to medium coverage (mean ~11.8×). The data set includes multiple individuals from eight representative cattle breeds, representing both the major taurine and indicine breeds used for both dairy and beef purposes. It provides unprecedented genome-wide resolution to interrogate CNV and a unique opportunity to fully explore the population-genetic properties and evolutionary contributions of multicopy genes related to breed-specific traits.

2. Material and methods

2.1. CNV calling, distribution, and association with other genomic features

We used a previously described segmentation algorithm to call CNVs.[16,48] Detailed sample selection and CNV calling methods can be found in Supplementary Material online. Association between CNVs and SDs was tested by Spearman's rank correlation using 100 kb windows as previously described.[49] Additional genomic features were obtained from public databases. Determination of the overlap between CNVRs and genomic features was performed as previously described.[10]

2.2. Population-genetic analyses and heatmap analysis

Inbreeding is a common feature in livestock due to selective mating and widespread use of artificial insemination. We filtered our samples based on known pedigrees constraining Wright's coefficient of relationship (r) to <0.25 to identify 69 unrelated individuals for the population-genetic analyses.

During the CNV discovery phase, a total of 1,148,528 windows of 1 kb were identified across the whole genome. Population-specific CNVs were estimated using the statistic V_{ST} developed by Redon et al.[21]. V_{ST} is calculated by considering $(V_T - V_S)/V_T$, where V_T is the variance in normalized copy numbers among all unrelated individuals and V_S is the average variance within each population, weighted for population size.

We next selected the top 1% diverse 1 kb windows ($n = 80$) from the distinct CNV regions to perform CNV genotyping using the partitioning around medoids (PAM) function in R.[50] A PAM procedure was used to cluster copy number ratios into discrete CN genotypes.[21] Similarly, we partitioned the copy numbers of each 1 kb window into three clusters, representing the low, mid, and high ranges and then coded them using the 0, 1, and 2 matrix for SNP genotyping. Population clustering was then performed using STRUCTURE v2.3.3,[51,52] assuming three ancestral populations ($k = 3$). This analysis between taurine and indicine cattle was initially run for values of the number of clusters (k) between 2 and 8. Each analysis was performed using 100,000 replicates and 100,000 burn-in cycles under admixture and correlated allele frequencies models. Reynolds' genetic distances among breeds were calculated using PHYLIP 3.69. To provide statistical support for the resulting clades, 10,000 bootstrap simulations were performed. The phylogenetic trees were visualized in FigTree 1.3.1 (http://tree.bio.ed.ac.uk/software/figtree/). Multidimensional scaling (MDS) analysis was conducted in PLINK 1.07 based on either CNV genotypes generated in this study or SNP genotypes generated using the same bovine HapMap populations as described previously.[53]

NimbleGen array CGH $\log_2$ ratios were calculated for each probe on a custom 2.1 million probe array for all animals. The reference animal, in all cases, was the Hereford cow used to generate the cattle reference assembly, L1 Dominette. All $\log_2$ ratio values that spanned the *GAT*, *GLYAT*, and *KRTAP9-1* genes were averaged across the gene's length for each animal individually. V_{ST} values contrasting taurine and indicine populations were calculated for these $\log_2$ ratios as previously described.[21]

Heatmaps were generated using the estimated CN windows for each animal as described previously.[16] The gplots (v 2.14.2) R package (http://cran.r-project.org/web/packages/gplots/index.html) was used to graph the CN values and generate heatmap representation of all lineage-specific gene duplications, deletions, and expansions identified in cattle breeds.

2.3. Gene analyses

Gene content of cattle CNVRs was assessed as previously described.[10] We performed DAVID analysis to test whether the terms were under- or overrepresented in CNVRs after Bonferroni correction.[10] We identified the lineage-specific or lineage-differential gene families using a heuristic approach with our 75 analysed individual animals. We divided the animals into breed, subspecies, and purpose groups as listed in Table 1 and used a weighted search algorithm to highlight CNVRs with a high tendency to exist solely within a specific group. The weighted search was accomplished as follows: for each CNVR we calculated a sum score that represented a hypothesis that the CNVR was unique to a specific breed/subspecies based on the animals that shared the CNVR. For each breed/subspecies/purpose group (G), we counted the number of animals (A) from G that shared the CNVR and imposed a penalty (P) for each animal that was not a member of the current G. The sets of G that were tested consisted of dairy, beef, Angus, Holstein, Limousin, Jersey, Romagnola, Nelore, Gir, Brahman, Taurus, and Indicus (membership was not mutually exclusive within the groups). The summed weight of A–P was calculated for each G, and if it exceeded a threshold of 3, it was selected as a putative subspecies/breed-specific or differential CNVR. We also employed a statistical method (V_{ST}) to identify copy number variable genes within our data set. Based on the gene CNs from each animal, we identified gene families that were stratified by subspecies differences using the statistic V_{ST}, as described above.

2.4. Haplotype network analysis

To explore the diversity of haplotypes and evolutionary relationships across populations, we retrieved the high-density SNP array data for these eight breeds generated by the Illumina BovineHD SNP Consortium as described previously.[53] Haplotypes and their frequencies were estimated separately for each breed using PHASE 2.1.[54,55] To obtain reliable results, we employed an iterative scheme to perform inference with 10,000 iterations and 10,000 burn-ins, also we increased the number of iterations of the final run of the algorithm using option -X100, for details see http://stephenslab.uchicago.edu/instruct2.1.pdf. Haplotype networks were constructed near functional genes such as *GAT/GLYAT*, *ASZ1*, *AOX1,* and *FZD3*. Phylogenetic relationships among the identified haplotypes were inferred through a median-joining network analysis using Network 4.6.12 (http://www.fluxus-engineering.com/).

2.5. Data release

All array CGH data have been submitted to the gene expression omnibus (http://www.ncbi.nlm.nih.gov/geo/) under the accession number GSE62990. Raw data and population genetic and evolutionary analysis results are available upon request for research purposes. Raw reads were deposited under the SRA Bioproject PRJNA277147 in SRA (http://www.ncbi.nlm.nih.gov/sra/).

3. Results and discussion

3.1. CNV discovery and experimental validations

After carefully excluding samples with low coverage in our initial survey, we focused on the remaining 75 individuals in our final data set (see Table 1 and Supplementary Table S1 for sample information and sequence coverage). We identified CNVs using a sliding window approach based on the previously published MrsFAST-WSSD method.[16,48] We discovered comparable average numbers of CNVs per individual across taurine (626.7) and indicine (591.2) cattle, suggesting our results based on the taurine reference assembly (UMD3.1) were not particularly biased against the indicine cattle. A full list of CNV calls (47,511) is presented in Supplementary Fig. S1 and Table S2. After merged across samples, these CNVs yielded 1,853 CNV regions (CNVR), which represent 87.5 Mb (3.1%) of the cattle genome (Supplementary Fig. S1 and Table S3). We then calculated absolute copy number values for 1 kb windows across the genome for each sequenced individual (see Materials and methods). As anticipated, the average normalized genome-wide copy number was 2.15 ± 0.1 for all copy number windows. We successfully performed 75 quantitative PCR (qPCR, Supplementary Table S4) and 25 array CGH experiments (Supplementary Fig. S2) to assess the false-positive discovery rate for our data set as previously described.[3,53] Detailed experimental validations can be found in Supplementary Material online. In summary, CNV calls made with sequence data were strongly correlated with array CGH data ($r^2 = 0.761$) and had an estimated 12% false-positive rate and a 19% false-negative rate based upon qPCR and array CGH, respectively.

We found that a large proportion of identified CNVRs (43.3%; 49.5 Mb) overlapped with the segmental duplication (SD) regions. We estimated pair-wise Spearman's rank correlations (Supplementary Table S5) of 0.084 and 0.098 for indicine and taurine CNVs and SD regions (both $P < 0.001$), which were similar to the previously reported human results.[49] A strong correlation of CNVs and SDs in cattle confirms that their formation mechanisms are mainly due to non-

Table 1. Samples and sequence data sets

Breed	Subspecies	Purpose	Animal count	Coverage range	CNV count	Average CNVs per animal[a]
Brahman (BRM)	*Bos t. indicus*	Beef	7	5–9×	3,836	548 (86)
Gir (GIR)	*Bos t. indicus*	Beef/dairy	6	5–14×	3,724	621 (30)
Nelore (NEL)	*Bos t. indicus*	Beef	8	6–20×	4,855	607 (38)
Angus (ANG)	*Bos t. taurus*	Beef	16	5–30×	11,657	729 (52)
Holstein (HOL)	*Bos t. taurus*	Dairy	22	4–20×	12,430	565 (80)
Jersey (JER)	*Bos t. taurus*	Dairy	6	4–13×	3,487	581 (46)
Limousin (LIM)	*Bos t. taurus*	Beef	6	5–10×	3,650	608 (48)
Romagnola (ROM)	*Bos t. taurus*	Beef/draft	4	6–10×	2,708	677 (20)

[a]Numbers in parentheses indicate 1 SD.

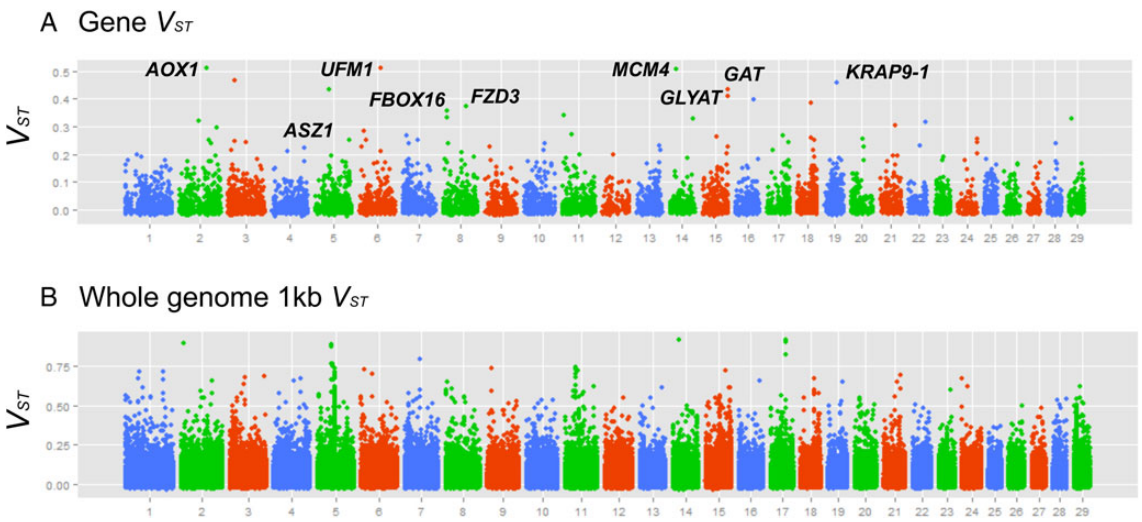

Figure 1. Population differentiation for copy number variation. Population differentiation, estimated by V_{ST}, is plotted along each chromosome for the two taurine and indicine comparisons: (A) RefSeq genes and (B) genome-wide 1 kb windows. Example CNVs exhibiting high population differentiation are labelled. This figure is available in black and white in print and in colour at *DNA Research* online.

allelic homologous recombination (NAHR).[10,56] In the following analyses, we mainly focused on the characterization of the high-confidence CNVs from autosomes.

3.2. Population genetics of cattle CNVs

To investigate the population genetics of CNVs, we first identified the frequencies of CNVRs in our data set (Supplementary Fig. S3). The average CNVR had a frequency of 29.3% (22 animals out of 75 total); however, the CNVR frequency tended towards a parabolic distribution with 835 CNVRs having a frequency ≤5% and 189 CNVRs having a frequency≥95% in our data set (Supplementary Fig. S3). As expected, rare events were often observed in only one subspecies/breed, whereas common CNVs (frequency >5%) were usually shared across subspecies/breeds.

To explore the population differentiation of CNVs between taurine and indicine cattle, we applied statistical measures of population differentiation using V_{ST}[21] to our dataset in three ways: (i) estimation of V_{ST} for genome-wide 1 kb CN windows; (ii) clustering of the top 1% of V_{ST} values for 1 kb CN windows; and (iii) estimation of V_{ST} using the average CN for annotated genes. Estimates of V_{ST} for all genome-wide CN windows and all CNVRs revealed a number of outliers with levels of population differentiation suggestive of population-specific selective pressures (Fig. 1). Among these outliers were CNVs near *AOX1*, *GAT/GLYAT*, *ASZ1*, *KRTAP9-1*, and *MCM4* (Fig. 1A).

We next selected the top 1% diverse 1 kb windows ($n = 80$)—identified by our V_{ST} calculations—from the distinct outlier CNVRs to perform CNV genotyping. The PAM algorithm is the most common implementation of *k*-medoid clustering, which is related to the *k*-means algorithm and the medoid shift algorithm.[50] Previously, a PAM algorithm was used to cluster copy number derived from array CGH $\log_2$ ratios into discrete CN genotypes.[21] Similarly, we partitioned the copy numbers of each 1 kb window into three discrete value clusters, representing low, mid, and high ranges and then coded them as 0, 1, and 2 values within a matrix for subsequent genotyping (Supplementary Table S6).

Using CNV genotype calls from the top 1% diverse 1 kb windows within CNVRs, we performed population clustering (Fig. 2). The proximity of an individual to each apex of the triangle indicates the proportion of that individual's genome that is estimated to have ancestry from each of the three inferred ancestral populations. The clustering of most indicine cattle (BRM, GIR, NEL) in the right bottom apex reveals the clear discrimination between indicine and taurine cattle. In contrast, the taurine cattle were scattered along the opposing side with the exception of ROM near the centre. This distribution of ROM individuals agreed with previous results based on SNP genotypes,[44] confirming that ROM has both taurine and indicine ancestries. It is also noted that ANG individuals were clustered together in the upper apex,

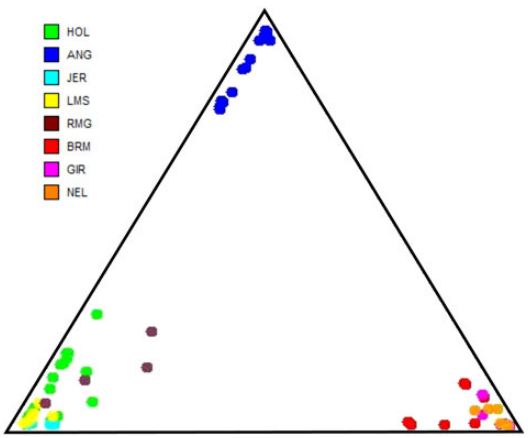

Figure 2. Population clustering based on CNV genotypes. A triangle plot showing the clustering of 69 lowly related cattle individuals assuming three ancestral populations ($k = 3$). The proximity of an individual to each apex of the triangle indicates the proportion of that genome that is estimated to have ancestry in each of the three inferred ancestral populations. The clustering together of most indicine individuals (BRM, GIR, NEL) in the right bottom apex indicates the clear discrimination between indicine and taurine cattle. In contrast, taurine cattle are scattered along the opposing side with the exception of ROM in the centre. ANG individuals were clustered together in the upper apex, while the other taurine cattle (HOL, LMS, JER) were dispersed around the left bottom corner, suggesting a possible discrimination between beef and dairy cattle. This figure is available in black and white in print and in colour at *DNA Research* online.

while the other taurine cattle (HOL, LMS, JER) were dispersed around the left bottom corner, suggesting a distinction between different taurine breeds. It is possible that these two clusters differentiate between continental European breeds of cattle from UK breeds or beef breeds from dairy breeds; however, it is also possible that our selected CNV markers may be subject to cryptic founder effects within our ANG individuals. We suspect that the addition of African cattle breeds to this dataset will better resolve the taurine cluster by providing a third distinct lineage. Still, we note that the clustering results were structurally similar to our results obtained with high-density SNP data derived from the same bovine HapMap samples (Supplementary Fig. S4) and other recently published results.[44] Based on the genotypes for these 80 loci (i.e. the top 1 kb windows) and using a neighbor-joining algorithm, we obtained a phylogenetic tree that generally agrees with the known cattle breed history (Supplementary Fig. S5). We also performed a MDS analysis based on CNV genotypes and compared it with the plot based on SNPs. Our plot confidently separated the indicine from the taurine cattle (Supplementary Fig. S6); however, the separation and clustering of the taurine cattle using CNVs were not superior to those based on SNPs, suggesting that CNV genotyping still has room for improvement.

Out of these 80 loci, 62 can be reliably assessed for their variable patterns and 54 of these loci, in turn, (87.10%, 54/62) are located in or near tandem duplications (Supplementary Table S6). This estimate was consistent with our initial genome-wide results that 90% of SD in cattle are tandem duplications in contrast to human and other primates, which show a preponderance of interspersed duplications.[56] This led us to speculate that while it is challenging to systematically genotype cattle duplication CNV events as shown by Genome STRiP results in human,[57] our relatively high cattle CNV genotyping accuracy is likely due to the vast majority of cattle CNV being tandem duplicates. Large tandem repeats or duplication CNVs in cattle could behave similarly like human tandem macrosatellites and multicopy genes. For these tandem duplications, it is likely that we made reasonable approximations of CNV genotyping calls by simply clustering the normalized copy numbers, as shown traditionally for macrosatellites and microsatellites.[58–61] Additionally, the tandem distribution pattern could contribute to the high LD at CNV loci as suggested previously,[62] thus the majority of CNV genotype calls could better represent local alleles. Combining these two factors, it is not surprising that our CNV-based results generally agree with SNP-based results. Of course, this hypothesis certainly warrants more investigation using larger sample sizes and other mammals like mouse and dog to further validate and improve CNV genotyping approaches.

To provide an evolutionary perspective to our analyses, we also created heatmaps using the CN values for regions within selected gene loci (Supplementary Fig. S7). These analyses of lineage-specific or lineage-differential CNVs separate subspecies/breeds into groupings that are consistent with the generally accepted cattle history.[44]

3.3. Gene analyses

We evaluated genes overlapped by cattle CNVs (Supplementary Table S7) and selected genes with known functions (Table 2). We observed an enrichment of CNVs intersecting with genes ($P < 0.0001$; Spearman's rank sum correlation), consistent with reduced evolutionary constraints acting on functionally redundant gene categories. We next used DAVID to identify basic biological functional categories for 361 genes overlapped by our identified CNVRs.[63] Like other mammals (human, mouse, and dog), statistically significant

Table 2. Selected copy number variable genes identified from population sequence data

Gene name	Function	Gene UMD3.1 coordinates	V_{ST}^{a}	Identified[b]
AOX1	Detoxification	chr2:89517708-89589232	0.5094	Hou, Bickhart, and this study
ASZ1	Spermatogenesis	chr4:51294534-51370343	0.2109	Only this study
CA1	Carbonic anhydrase	chr14:79520632-79530892	0.3270	Hou, Bickhart, and this study
CFH	Complement factor	chr16:5486704-6172566	0.0483	Hou, Bickhart, and this study
DDX21	Translation initiation	chr28:25376358-25399769	0.2375	Only this study
DENR	Translation initiation	chr29:7723699-7725004	0.3285	Only this study
FBXO16	Ubiquitin protein ligase	chr8:10095869-10128675	0.3558	Bickhart and this study
FZD3	Nervous system	chr8:10002971-10091175	0.3334	Bickhart and this study
GAL3ST1	Glycolipid catalysis	chr17:71660016-71678806	0.2435	Hou and this study
GAT	Detoxification	chr15:83472190-83493607	0.4336	Liu, Hou, Bickhart, and this study
GLYAT	Detoxification	chr15:83455512-83469280	0.4083	Liu, Hou, Bickhart, and this study
GLYATL2	Biological oxidation	chr15:83508339-83515102	0.2257	Bickhart and this study
KRTAP9-1	Keratin family	chr19:42101853-42103421	0.4578	Bickhart and this study
LMBRD2	Function unknown	chr20:38116509-38163145	0.2573	Only this study
PGR	Progesterone receptor	chr15:8207682-8222806	0.0103	Only this study
PNPLA2	Adipose tissue regulation	chr29:50742384-50747161	0.0000	Only this study
PRG3	Carbohydrate binding	chr15:81920283-81926082	0.2041	Liu, Bickhart, and this study
RAET1G/ULBP17	MHC class 1 related	chr9:88231932-88402262	0.0000	Liu, Hou, Bickhart, and this study
RICTOR	Cell growth	chr20:35376523-35514753	0.1048	Only this study
SEC23A	Vesicle transport	chr21:49489514-49555507	0.3050	Only this study
SERPINB4	Protease inhibitor	chr24:62364701-62371668	0.2418	Liu, Hou, Bickhart, and this study
SUB1	Transcriptional activation	chr20:41122022-41143914	0.2282	Only this study
TMED2	Secretory vesicle transport	chr17:54330420-54338333	0.0000	Only this study
UFM1	Ubiquitin	chr6:71051155-71053533	0.5121	Only this study
ZNF280B	Negative regulation of p53	chr17:51251538-51262528	0.2658	Liu, Hou, and this study

[a] V_{ST} was calculated from the comparison between the taurine and indicine individuals.

[b] Liu, Hou, and Bickhart: we focused on the comparisons with the published CNV results based on the same bovine HapMap samples using array CGH,[10] BovineHD SNP array,[53] and individual NGS,[16] respectively.

overrepresentations were observed for multiple categories including chromosome maintenance, immunity and cytoskeleton components (Supplementary Table S8). We then studied how variable genes were distributed across subspecies/breeds using either a heuristic approach based on CNV presence/absence or gene CN per individual (V_{ST}).

3.4. Lineage-specific CNV genes based on a heuristic approach

We first identified lineage specific, copy number variable genes (CNV genes) using a heuristic approach (see Materials and methods). Dairy cattle-specific CNVs tended to be present at low frequencies in our 28 dairy cattle, and they manifested as small copy number changes of affected genes. Several of these dairy-specific CNVRs were found to intersect genes related to cellular growth and development pathways, including *RICTOR* (rapamycin-insensitive companion of mTOR)[64] and *TMED2*.[65] We also identified several lipid metabolism genes that overlapped CNVRs exclusive to beef cattle (over 40 samples). Within our Angus data set, we discovered six animals that had a predicted heterozygous duplication of *PNPLA3* (the Patatin-like phospholipase domain-containing protein 3). This gene is expressed in adipose tissue and liver, and is associated with the *de novo* synthesis of fatty acids.[66] In indicine beef cattle (Nelore, Brahman, and Gir), we also detected a duplication in a predicted Ensembl gene (ENSBTAT00000043749) containing functional domains related to lipid metabolism.

3.5. Gene family expansion, diversity, and evolution

We identified copy number variable genes among the different subspecies/breeds using V_{ST} statistics. We defined highly stratified genes as genes having V_{ST} values >0.2. CN plots for these stratified genes showed clear differences in the average CN value for taurine and indicine animals (Fig. 3A). Based on V_{ST} values, *ZNF280B*, *FBXO16*, *KRTAP9-1*, *MCM4*, *SERPINB4*, *CA1*, *FZD3*, *GLAYAT*, *GAT*, *MANBA*, and *DENR* were the most stratified genes. To provide orthogonal experimental support for the sequence-based V_{ST} results, we also retrieved $\log_2$ ratios from array CGH data for the same animals. Representative results are shown in Supplementary Fig. S8 for the *GAT*, *GLYAT,* and *KRTAP9-1* genes, further confirming the sequence-based V_{ST} results.

CN stratified genes tended to be immune system related, which is expected given the different environmental challenges in the history of evolution of taurine and indicine cattle. One of these stratified CNVs represents a significantly higher duplication of the *KRTAP9-1* gene in taurine cattle, which is a paralog of *KRTAP9-2* that was previously reported to likely be involved in indicine tick resistance (Fig. 3B). A duplication of the *MCM4* gene (Fig. 3C) was found in indicine cattle compared with taurine cattle. We also confirmed several other gene families appearing to be copy number variable, including lysozyme, defensin, and unique long binding protein (*ULBP*) families and the major histocompatibility complex (*MHC*).

Discoveries of high-frequency gene duplications suggest that the affected gene families are currently expanding in a 'gene family birth and death' model as described by Nei and Rooney.[67] Such multicopy genes, if present in a sufficiently large proportion of the population, can be thought of as signs of diversifying selection or selection by overdominance.[68] One example of this can be found in the Olfactory Receptor (OR) gene family, which has several member genes that detect odorant molecules through combinatorial binding across other paralogous family members.[69] Therefore, the duplication and subsequent

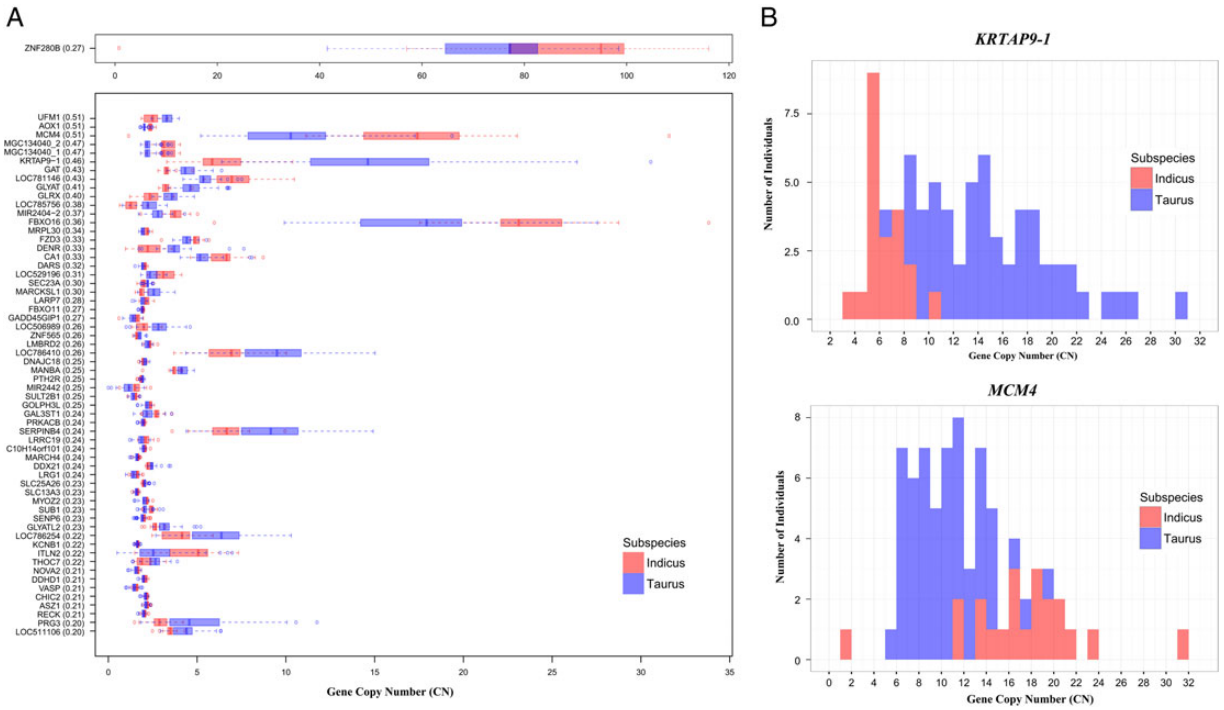

Figure 3. Cattle gene family copy number diversity and evolution. The genes most stratified by copy number on the basis of V_{ST} analysis of taurine and indicine cattle (A). The most copy number variable genes in both taurine and indicine subspecies (legend insets denote group colors) tended to be immune system-related genes. Histograms showing the distributions of copy numbers among the unrelated individuals in each group are plotted for the *KRTAP9-1* gene (B) and the *MCM4* gene (C). *X*-axis values indicate copy number and *Y*-axis values indicate sample count. Individual copy number values for each gene can be found in Supplementary Table S7. This figure is available in black and white in print and in colour at *DNA Research* online.

mutation of OR gene members allow for a greater range of odorant detection for a host organism. Indeed, out of 134 annotated OR genes, we have identified 31 (23.1%) individual genes that have predicted duplications in our data set.

We have detected several additional gene families that appear to be subject to a high degree of duplication in our data set, and these families likely represent classes of genes that are in the processes of subfunctionalization and neofunctionalization in cattle. They include a cluster of prolactin-related protein family (*PRP*) genes that appears to be duplicated in 96% (74/75) animals. It was previously discovered using the BovineSNP50 array[70]; however, we refined the event from 2.4 down to 0.7 megabases in size. Another locus containing several pregnancy-associated glycoprotein (*PAG*) family members was found to be duplicated in all animals within CNVR 1717.

3.6. Haplotype network analyses near selected multicopy genes

It is important to note that CNVs in some loci may have different alleles. Earlier results also suggest that the diversity of a subset of multicopy genes like human OR genes may have been maintained by balancing selection, in the form of overdominance.[68] For example, a 660 kb deletion with antagonistic effects on fertility and milk production was recently found at high frequency in Nordic Red cattle, providing evidence for balancing selection of CNVs in livestock.[71] To investigate the potential effect of overdominance on selection and evolution of multicopy genes, we further investigated haplotype evolution pattern using the BovineHD SNP array. We obtained 11 haplotypes within the 50.3 kb haploblock region near the *GLYAT/GAT* locus (Fig. 4A). The most common haplotype, H1 (with frequency of 70.06%), was mainly found in taurine cattle (HOL, ANG, JER, and LMS) and only minor portions were found in indicine cattle (ROM, BRM, GIR, and NEL) (Fig. 4A). H2 (with frequency of 10.56%) included a large proportion of taurine cattle (ANG, LMS, and ROM). We also observed two haplotypes exclusive to indicine cattle with a combined frequency of 10.54%: H3 and H4. Altogether, this pattern indicated that separate haplotypes were clustered only for the indicine cattle (BRM, GIR, and NEL), while other common haplotypes were identified for taurine cattle (HOL, ANG, LMS, JER, and ROM). BRM and ROM were the only exceptions, as they were often associated with both taurine and indicine cattle. This was not unexpected, as it mirrors the complex ancestral backgrounds of these two breeds, since BRM cattle are a known indicine breed with taurine influence and ROM share distinctive genetic ancestry with indicine cattle.[44] We also found similar results for other copy number variable genes, such as *ASZ1* (Fig. 4B), *AOX1*, and *FZD3* (Supplementary Figs S9 and S10). These haplotype network analyses suggest that for a subset of multicopy genes: (i) common overlapping allelic haplotypes were often present within the taurine cattle, while separate distinct haplotypes were present in the indicine cattle, suggesting different evolutionary history for these two cattle subspecies; and (ii) there was high allelic diversity near multicopy genes maintained by balancing selection, in the form of overdominance, suggesting that they have been under different selection pressures in these two cattle subspecies.

3.7. The impacts of the reference genome assembly

Different versions of cattle reference genome assemblies (Btau_4.0 and UMD3.1) have different RefSeq gene annotations, particularly in CN variable regions of the cattle genome. For example, the *CATHL4* gene, for which we previously reported copy number change,[16] was located on chrUn of UMD3.1. Since our RD method uses a window approach that relies on large genomic segments, this prevented us from assessing the copy number status of *CATHL4* on chrUn. However, we did find *CATHL1* was copy number variable in this study. Similarly, we were unable to assess the copy number of *KRTAP9-2*; however, we detected copy number changes for one of its paralogs, *KRTAP9-1*, in this study.

3.8. Limitations and future directions

Similar to SNP and microsatellite, CNV distribution within and among populations seems to be shaped by mutation, recombination, gene conversion, selection, and demographic history.[18,19,28,72] However, CNV genetic markers may not be currently compatible with current population analyses, because CNVs violate the classical population genetics assumptions based on the infinite allele model and the infinite site model for SNP. Compared with SNPs, limitations of CNVs as markers were observed in this study probably due to their distinct mutation mechanisms, high mutation rates, heterogeneities among loci, and uncertainties related to allele calling. Similar observations were also reported for microsatellite—primarily due to similar limitations of detection and variability.[73,74] For example, although NAHR is believed to be responsible for most of large duplication CNVs in cattle, inference, and predictions on the forces influencing populations require the modelling of the mutational process generating CNV. We currently lack such a model for the large duplications present in cattle. This is also compounded by the fact that homoplasy

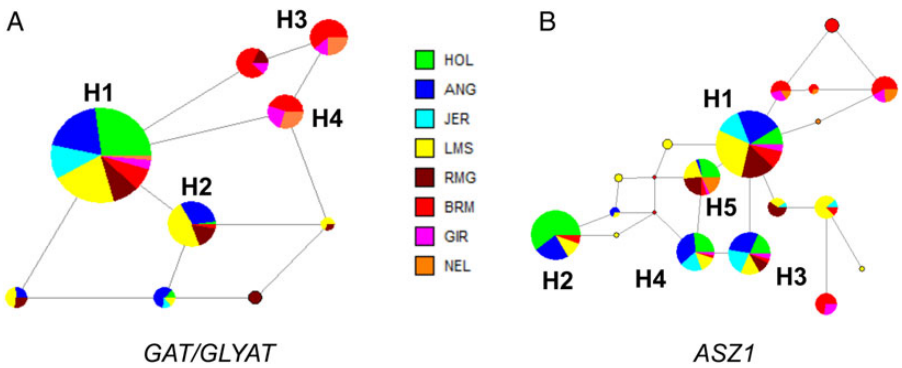

Figure 4. Haplotype networks of two loci. (A) The GAT/GLYAT locus and (B) the ASZ1 locus. Each node represents a different haplotype, with the size of the circle proportional to frequency. Circles are colour coded according to breeds. This figure is available in black and white in print and in colour at *DNA Research* online.

caused by recurrent events is expected to occur relatively often for CNV compared with SNP because of their high mutation rates.

Despite the aforementioned limitations, this study represents one of the first attempts to genotype CNVs within large, diverse cattle populations using sequence data. Although beyond the scope of this study, a comparison with human-centric CNV genotyping methods using cattle sequence data will provide a useful contrast in approaches. Our results provide a new glimpse into the diversity of selective pressures during cattle speciation. We confirmed that cattle are strikingly diverse, despite relatively low estimated current population sizes for several taurine cattle as shown previously.[37] Our population-genetic analyses based on CNVs reveal the population structures of these taurine and indicine cattle and uncover hundreds of CNVs showing elevated population differentiation near important functional genes. We highlighted several subspecies specific or differential CNV gene overlaps that are likely subject to subfunctionalization and neofunctionalization. We also identified key regions of the cattle genome that are subject to variation and reported several potential genes affecting productive traits. These discoveries provide a basis for future efforts to genotype and track large CNVs in cattle. More sequencing data from the 1000 Bull Genomes Project[75] or the analysis of additional outlier groups (e.g. African cattle breeds) will help to validate and refine the link between genomic copy number in these regions or different alleles with production and health traits.

Authors' contributions

G.E.L. and D.M.B. conceived and designed the experiments. D.M.B., L.X., J.L.H., J.B.C., and J.S. performed *in silico* prediction and computational analyses. D.M.B. and L.X. performed aCGH, qPCR confirmation. D.J.N., S.G.S., J.F.G., C.P.V.T., T.S.S., R.D.S., J.F.T., and H.A.L. collected samples and generated sequence data. G.E.L. and D.M.B. wrote the paper.

Acknowledgements

We thank members of the Illumina BovineHD SNP Consortium for sharing their samples and data. We would thank the Council on Dairy Cattle Breeding (Reynoldsburg, OH) for access to animal phenotypic information, and we thank the Cooperative Dairy DNA Repository (Beltsville, MD) for access to semen stocks used to generate Illumina DNA sequence libraries. We also thank T. Brown, R. Anderson, and A. Dimtchev for technical assistance. Mention of trade names or commercial products in this article is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the US Department of Agriculture.

Supplementary data

Supplementary data are available at www.dnaresearch.oxfordjournals.org.

Funding

This work was supported in part by Agriculture and Food Research Initiative (AFRI) competitive grant No. 2011-67015-30183 from the USDA National Institute of Food and Agriculture (NIFA) Animal Genome Program. J.F.T. and C. P.V.T. were supported by AFRI competitive grant No. 2009-65205-05635 from the USDA NIFA. J.F.T. and R.D.S. were further supported by AFRI competitive grants No. 2011-68004-30367, 2011-68004-30214 and 2013-68004-20364 from the USDA NIFA Animal Genome Program. Funding to pay the Open Access publication charges for this article was provided by the USDA National Institute of Food and Agriculture (NIFA) Animal Genome Program.

Conflict of interest statement

T.S.S. is an employee of Acceligen Inc. of Animal Agriculture Subsidiary of Recombinetics, Inc. All other authors declare no potential conflict of interest.

References

1. Mills, R.E., Walter, K., Stewart, C., et al. 2011, Mapping copy number variation by population-scale genome sequencing, *Nature*, **470**, 59–65.
2. Sudmant, P.H., Rausch, T., Gardner, E.J., et al. 2015, An integrated map of structural variation in 2,504 human genomes, *Nature*, **526**, 75–81.
3. Sudmant, P.H., Kitzman, J.O., Antonacci, F., et al. 2010, Diversity of human copy number variation and multicopy genes, *Science*, **330**, 641–6.
4. Girirajan, S., Campbell, C.D. and Eichler, E.E. 2011, Human copy number variation and complex genetic disease, *Annu. Rev. Genet.*, **45**, 203–26.
5. Bae, J.S., Cheong, H.S., Kim, L.H., et al. 2010, Identification of copy number variations and common deletion polymorphisms in cattle, *BMC Genomics*, **11**, 232.
6. Chen, W.K., Swartz, J.D., Rush, L.J. and Alvarez, C.E. 2009, Mapping DNA structural variation in dogs, *Genome Res.*, **19**, 500–9.
7. Fadista, J., Thomsen, B., Holm, L.E. and Bendixen, C. 2010, Copy number variation in the bovine genome, *BMC Genomics*, **11**, 284.
8. Fontanesi, L., Beretti, F., Riggio, V., et al. 2009, Copy number variation and missense mutations of the agouti signaling protein (ASIP) gene in goat breeds with different coat colors, *Cytogenet. Genome Res.*, **126**, 333–47.
9. Fontanesi, L., Beretti, F., Martelli, P.L., et al. 2011, A first comparative map of copy number variations in the sheep genome, *Genomics*, **97**, 158–65.
10. Liu, G.E., Hou, Y., Zhu, B., et al. 2010, Analysis of copy number variations among diverse cattle breeds, *Genome Res.*, **20**, 693–703.
11. Nicholas, T.J., Cheng, Z., Ventura, M., Mealey, K., Eichler, E.E. and Akey, J.M. 2009, The genomic architecture of segmental duplications and associated copy number variants in dogs, *Genome Res.*, **19**, 491–9.
12. Ramayo-Caldas, Y., Castello, A., Pena, R.N., et al. 2010, Copy number variation in the porcine genome inferred from a 60 k SNP BeadChip, *BMC Genomics*, **11**, 593.
13. Kijas, J.W., Barendse, W., Barris, W., et al. 2011, Analysis of copy number variants in the cattle genome, *Gene*, **482**, 73–7.
14. Snyder, M., Du, J. and Gerstein, M. 2010, Personal genome sequencing: current approaches and challenges, *Genes Dev.*, **24**, 423–31.
15. Rubin, C.J., Zody, M.C., Eriksson, J., et al. 2010, Whole-genome resequencing reveals loci under selection during chicken domestication, *Nature*, **464**, 587–91.
16. Bickhart, D.M., Hou, Y., Schroeder, S.G., et al. 2012, Copy number variation of individual cattle genomes using next-generation sequencing, *Genome Res.*, **22**, 778–90.
17. Durkin, K., Coppieters, W., Drogemuller, C., et al. 2012, Serial translocation by means of circular intermediates underlies colour sidedness in cattle, *Nature*, **482**, 81–4.
18. Conrad, D.F. and Hurles, M.E. 2007, The population genetics of structural variation, *Nat. Genet*, **39**, S30–6.
19. Kato, M., Kawaguchi, T., Ishikawa, S., et al. 2010, Population-genetic nature of copy number variations in the human genome, *Hum. Mol. Genet.*, **19**, 761–73.
20. Sudmant, P.H., Mallick, S., Nelson, B.J., et al. 2015, Global diversity, population stratification, and selection of human copy-number variation, *Science*, **349**, aab3761.
21. Redon, R., Ishikawa, S., Fitch, K.R., et al. 2006, Global variation in copy number in the human genome, *Nature*, **444**, 444–54.
22. Conrad, D.F., Pinto, D., Redon, R., et al. 2009, Origins and functional impact of copy number variation in the human genome, *Nature*, **464**, 704–12.
23. Jakobsson, M., Scholz, S.W., Scheet, P., et al. 2008, Genotype, haplotype and copy-number variation in worldwide human populations, *Nature*, **451**, 998–1003.
24. Iskow, R.C., Gokcumen, O. and Lee, C. 2012, Exploring the role of copy number variants in human adaptation, *Trends Genet.*, **28**, 245–57.

25. Zhang, F., Gu, W., Hurles, M.E. and Lupski, J.R. 2009, Copy number variation in human health, disease, and evolution, *Annu. Rev. Genomics Hum. Genet.*, **10**, 451–81.
26. Perry, G.H., Yang, F., Marques-Bonet, T., et al. 2008, Copy number variation and evolution in humans and chimpanzees, *Genome Res.*, **18**, 1698–710.
27. Paudel, Y., Madsen, O., Megens, H.J., et al. 2013, Evolutionary dynamics of copy number variation in pig genomes in the context of adaptation and domestication, *BMC Genomics*, **14**, 449.
28. Brown, K.H., Dobrinski, K.P., Lee, A.S., et al. 2012, Extensive genetic diversity and substructuring among zebrafish strains revealed through copy number variant analysis, *Proc. Natl Acad. Sci. USA*, **109**, 529–34.
29. Berglund, J., Nevalainen, E.M., Molin, A.M., et al. 2012, Novel origins of copy number variation in the dog genome, *Genome Biol.*, **13**, R73.
30. Perry, G.H., Dominy, N.J., Claw, K.G., et al. 2007, Diet and the evolution of human amylase gene copy number variation, *Nat. Genet.*, **39**, 1256–60.
31. Kidd, J.M., Newman, T.L., Tuzun, E., Kaul, R. and Eichler, E.E. 2007, Population stratification of a common APOBEC gene deletion polymorphism, *PLoS Genet.*, **3**, e63.
32. Xue, Y., Sun, D., Daly, A., et al. 2008, Adaptive evolution of UGT2B17 copy-number variation, *Am. J. Hum. Genet.*, **83**, 337–46.
33. Polimanti, R., Piacentini, S., Iorio, A., et al. 2015, Haplotype differences for copy number variants in the 22q11.23 region among human populations: a pigmentation-based model for selective pressure, *Eur. J. Hum. Genet.*, **23**, 116–23.
34. McCarroll, S.A., Hadnott, T.N., Perry, G.H., et al. 2006, Common deletion polymorphisms in the human genome, *Nat. Genet.*, **38**, 86–92.
35. Xu, L., Cole, J.B., Bickhart, D.M., et al. 2014, Genome wide CNV analysis reveals additional variants associated with milk production traits in Holsteins, *BMC Genomics*, **15**, 683.
36. Bradley, D.G., Machugh, D.E., Cunningham, P. and Loftus, R.T. 1996, Mitochondrial diversity and the origins of African and European cattle, *Proc. Natl Acad. Sci. USA*, **93**, 5131–5.
37. The Bovine HapMap Consortium 2009, Genome wide survey of SNP variation uncovers the genetic structure of cattle breeds, *Science*, **324**, 528–32.
38. MacHugh, D.E., Shriver, M.D., Loftus, R.T., Cunningham, P. and Bradley, D.G. 1997, Microsatellite DNA variation and the evolution, domestication and phylogeography of taurine and zebu cattle (*Bos taurus* and *Bos indicus*), *Genetics*, **146**, 1071–86.
39. Hiendleder, S., Lewalski, H. and Janke, A. 2008, Complete mitochondrial genomes of *Bos taurus* and *Bos indicus* provide new insights into intra-species variation, taxonomy and domestication, *Cytogenet. Genome Res.*, **120**, 150–6.
40. Loftus, R.T., MacHugh, D.E., Bradley, D.G., Sharp, P.M. and Cunningham, P. 1994, Evidence for two independent domestications of cattle, *Proc. Natl Acad. Sci. USA*, **91**, 2757–61.
41. Larson, G., Piperno, D.R., Allaby, R.G., et al. 2014, Current perspectives and the future of domestication studies, *Proc. Natl Acad. Sci. USA*, **111**, 6139–46.
42. Troy, C.S., MacHugh, D.E., Bailey, J.F., et al. 2001, Genetic evidence for Near-Eastern origins of European cattle, *Nature*, **410**, 1088–91.
43. Hanotte, O., Bradley, D.G., Ochieng, J.W., Verjee, Y., Hill, E.W. and Rege, J.E. 2002, African pastoralism: genetic imprints of origins and migrations, *Science*, **296**, 336–9.
44. Decker, J.E., McKay, S.D., Rolf, M.M., et al. 2014, Worldwide patterns of ancestry, divergence, and admixture in domesticated cattle, *PLoS Genet.*, **10**, e1004254.
45. Andersson, L. and Georges, M. 2004, Domestic-animal genomics: deciphering the genetics of complex traits, *Nat. Rev. Genet.*, **5**, 202–12.
46. VanRaden, P.M., Van Tassell, C.P., Wiggans, G.R., et al. 2009, Invited review: reliability of genomic predictions for North American Holstein bulls, *J. Dairy Sci.*, **92**, 16–24.
47. Porto-Neto, L.R., Sonstegard, T.S., Liu, G.E., et al. 2013, Genomic divergence of zebu and taurine cattle identified through high-density SNP genotyping, *BMC Genomics*, **14**, 876.
48. Alkan, C., Kidd, J.M., Marques-Bonet, T., et al. 2009, Personalized copy number and segmental duplication maps using next-generation sequencing, *Nat. Genet.*, **41**, 1061–7.
49. Kim, P.M., Lam, H.Y., Urban, A.E., et al. 2008, Analysis of copy number variants and segmental duplications in the human genome: evidence for a change in the process of formation in recent evolutionary history, *Genome Res.*, **18**, 1865–74.
50. Reynolds, A.P., Richards, G., de la Iglesia, B. and Rayward-Smith, V.J. 2006, Clustering rules: a comparison of partitioning and hierarchical clustering algorithms, *J. Math. Model. Algor.*, **5**, 475–504.
51. Falush, D., Stephens, M. and Pritchard, J.K. 2003, Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies, *Genetics*, **164**, 1567–87.
52. Pritchard, J.K., Stephens, M. and Donnelly, P. 2000, Inference of population structure using multilocus genotype data, *Genetics*, **155**, 945–59.
53. Hou, Y., Bickhart, D.M., Hvinden, M.L., et al. 2012, Fine mapping of copy number variations on two cattle genome assemblies using high density SNP array, *BMC Genomics*, **13**, 376.
54. Stephens, M., Smith, N.J. and Donnelly, P. 2001, A new statistical method for haplotype reconstruction from population data, *Am. J. Hum. Genet.*, **68**, 978–89.
55. Crawford, D.C., Bhangale, T., Li, N., et al. 2004, Evidence for substantial fine-scale variation in recombination rates across the human genome, *Nat. Genet.*, **36**, 700–6.
56. Liu, G.E., Ventura, M., Cellamare, A., et al. 2009, Analysis of recent segmental duplications in the bovine genome, *BMC Genomics*, **10**, 571.
57. Handsaker, R.E., Van Doren, V., Berman, J.R., et al. 2015, Large multiallelic copy number variations in humans, *Nat. Genet.*, **47**, 296–303.
58. Brahmachary, M., Guilmatre, A., Quilez, J., et al. 2014, Digital genotyping of macrosatellites and multicopy genes reveals novel biological functions associated with copy number variation of large tandem repeats, *PLoS Genet.*, **10**, e1004418.
59. Willems, T., Gymrek, M., Highnam, G., Mittelman, D. and Erlich, Y. 2014, The landscape of human STR variation, *Genome Res.*, **24**, 1894–904.
60. Carlson, K.D., Sudmant, P.H., Press, M.O., Eichler, E.E., Shendure, J. and Queitsch, C. 2015, MIPSTR: a method for multiplex genotyping of germline and somatic STR variation across many individuals, *Genome Res.*, **25**, 750–61.
61. Fungtammasan, A., Ananda, G., Hile, S.E., et al. 2015, Accurate typing of short tandem repeats from genome-wide sequencing data and its applications, *Genome Res.*, **25**, 736–49.
62. Schrider, D.R. and Hahn, M.W. 2010, Lower linkage disequilibrium at CNVs is due to both recurrent mutation and transposing duplications, *Mol. Biol. Evol.*, **27**, 103–11.
63. Huang, D.W., Sherman, B.T. and Lempicki, R. 2009, Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources, *Nat. Protoc.*, **4**, 44–57.
64. Jacinto, E., Loewith, R., Schmidt, A., et al. 2004, Mammalian TOR complex 2 controls the actin cytoskeleton and is rapamycin insensitive, *Nat. Cell Biol.*, **6**, 1122–8.
65. Jerome-Majewska, L.A., Achkar, T., Luo, L., Lupu, F. and Lacy, E. 2010, The trafficking protein Tmed2/p24β(1) is required for morphogenesis of the mouse embryo and placenta, *Dev. Biol.*, **341**, 154–66.
66. Li, J.Z., Huang, Y., Karaman, R., et al. 2012, Chronic overexpression of PNPLA3I148M in mouse liver causes hepatic steatosis, *J. Clin. Invest.*, **122**, 4130–44.
67. Nei, M. and Rooney, A.P. 2005, Concerted and birth-and-death evolution of multigene families, *Annu. Rev. Genet.*, **39**, 121–52.
68. Alonso, S., Lopez, S., Izagirre, N. and de la Rúa, C. 2008, Overdominance in the human genome and olfactory receptor activity, *Mol. Biol. Evol.*, **25**, 997–1001.
69. Malnic, B., Hirono, J., Sato, T. and Buck, L.B. 1999, Combinatorial receptor codes for odors, *Cell*, **96**, 713–23.

70. Cicconardi, F., Chillemi, G., Tramontano, A., et al. 2013, Massive screening of copy number population-scale variation in Bos taurus genome, *BMC Genomics*, **14**, 124.
71. Kadri, N.K., Sahana, G., Charlier, C., et al. 2014, A 660-Kb deletion with antagonistic effects on fertility and milk production segregates at high frequency in Nordic Red cattle: additional evidence for the common occurrence of balancing selection in livestock, *PLoS Genet.*, **10**, e1004049.
72. Teshima, K.M. and Innan, H. 2012, The coalescent with selection on copy number variants, *Genetics*, **190**, 1077–86.
73. Haasl, R.J. and Payseur, B.A. 2011, Multi-locus inference of population structure: a comparison between single nucleotide polymorphisms and microsatellites, *Heredity (Edinb.)*, **106**, 158–71.
74. Putman, A.I. and Carbone, I. 2014, Challenges in analysis and interpretation of microsatellite data for population genetic studies, *Ecol. Evol.*, **4**, 4399–428.
75. Daetwyler, H.D., Capitan, A., Pausch, H., et al. 2014, Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle, *Nat. Genet.*, **46**, 858–65.